# A Survey on Learning Crowdsourced User Preferences for Visual Summarization of Image Collections

**Rupali Tanaji Waghmode[1], Nikita J. Kulkarni[2]**

[1]Pune University, Dnyanganga College of Engineering and Research, Pune-411046, India

[2]Professor, Pune University, Dnyanganga College of Engineering and Research, Pune-411046, India

**Abstract:** ***We present a new approach for selecting images which are suitable for inclusion in the visual summaries. This approach is designed on the basis of how people generally think of summarizing image collections. For obtaining large number of manually created visual image summaries and criteria which guide user's for selection of images we use Amazon Mechanical Turk Crowd sourcing platform. This technique utilizes the content and context of images, image popularities, similarities between images, sentimental analysis. We describe images not only on the basis of their properties but also we consider the fact behind images that are related semantically. This increases efficiency and enables aesthetic appeal, proliferation of sentiment, and various emotions associated with a particular group of images. We examine the trend of a low inter-user contract, which is helping to make a computerized evaluation of aesthetic summaries and propose a solution influenced by the text summarization and machine interpretation communities. The studies conducted on a collection of geo-referenced Flickr image collections demonstrate the potency of our image selection approach.***

**Keywords:** Visual, Crowdsourcing, aesthetic, user-informed image selection, set evaluation

## 1. Introduction

Very quick growth of the quantity of digital multimedia data available in professional and personal collections along with the social networking and content sharing websites has established the necessity of powerful tools enabling representation, summarization, analysis and abstraction of data for more feasible and effective retrieval and browsing. Summarization techniques, specifically, aim at providing a concise representation of an individual multimedia data collection or data document. Based on the kind of the application and data domain, summaries may contain videos, segments, text, images.

We take the problem of generation visual summaries of geographical places as trial use case to show the benefits of the proposed user-informed picture selection concept. The paper makes the following principal benefits:

- This topic present a new approach based on how humans select images for visual summaries, which was collected with a large-scale crowd sourcing study, as the basis for a novel method for automatically selecting images for visual summarization.
- This approach uses extracted features of images and RankSVM method to generate a list of images ranked by their suitability for inclusion in a visual summary.
- The selected set of pictures can be used as a "general purpose" visual summary or as a starting point in building a summary with particular properties.

## 2. Related work

### 2.1 Visual Summarization

In general, visual summarization intends to give compact information of a single video, set of videos or an image collection.

In [2] we conjecture image analysis, tag data and images' explicit and implicit metadata to extract meaningful features from community-contributed datasets. We use tags i.e. text labels associated with images by users and location metadata to detect tags and location that represent landmark or geographic features. We perform visual analysis of images associated with discovered landmarks to extract representative sets of images for each landmark. We cluster the landmark images into visually similar groups by using various image processing methods, as well as generate links between those images that contain the same visual objects. Based on the clustering and on the generated link structure, we recognize canonical views, as well as select the top representative images for each such view. This technique helps for getting diverse and representative results for landmark searches. One of the demerit is focuses on best views of the landmark itself.

In [3] we present a framework of summarizing tourist destinations by leveraging the rich textual and visual information in large amount of user-generated travelogues and photos on the Web. The framework first discovers location representative tags from travelogues and then selects relevant and representative photos to visualize these tags. The learnt tags and selected photos are finally organized appropriately to provide an informative summary which describes a given destination both textually and visually. Experimental results based on a large collection of travelogues and photos show promising results on destination summarization.

In [4] we present a new technique for automatic visual summarization of geographic area. We present a new retrieval technique and learning framework for automatic visual summarization of geographic area. Here we take geo-coordinates of particular location as input and then download images within set radius from Flickr website. It uses metadata, textual and visual modalities of images. Then we represent semantic relations between images based on user interaction. In this technique System uses Multimodal Image Context Graph (ICG) which combines visual, textual and other modalities together. The theory of random walks i.e. RWR over graph is used to compute representative score (RS) and diversity score (DS). This new method does not require input from human.

### 2.2 Summary Evaluation

The Recent research has investigated different types of summaries, methods to create and methods to evaluate them. In [5] The Bilingual Evaluation Understudy (BLUE) is method for automatically evaluating the quality of machine translation based on ngram co-occurrence scoring. It is now the new scoring measure used in the NIST (NIST 2002) translation benchmarks. The main idea of BLEU is to measure the similarities between a candidate translation and a set of reference translations. BLEU compares a candidate translation with several human-generated reference translations using n-gram cooccurrence statistics.

In [6] ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure is used for text summarization evaluation. This metric calculates the number of overlapping units between the summary candidates generated by computer and several ground truth summaries built by humans. Several variants of the metric are introduced, such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. To calculate the effectiveness of ROUGE measures, we calculate the correlation between ROUGE assigned summary scores and human assigned summary scores.

In [7] we study VERT-Video Evaluation by Relevant Threshold is the algorithm to automatically evaluate the quality of video summaries. We use ideas from ROUGE and BLEU and extend these measures to the domain of video summarization. Our approach focus on the selection of relevant key frames, as a video skim can be easily constructed by concatenating video clips extracted around the selected key frames. Here we believe that the temporal order of key frames is not as important as the word order in a sentence, we rather use the key frame importance rank in the selection. We select a set of video sequences $v_{1,}$ $v_{2,}$ $v_{3,}$ $v_k$ related to a given topic. These sequences are divided into shots or subshots, and each shot is represented by one or more keyframes. Based on shots, subshots or keyframes, a selection of the video content to be included in the summary is performed. This selection may be ordered, with the most important content being selected first. The selected content is assembled into a video summary, either in the form of a photo album or a video skim. VERT metric compares a set of computer selected key frames with several reference sets of human-selected key frames. Since BLEU is precision measure and ROUGE is recall measure, we propose VERT-Precision and VERT-Recall respectively.

### 2.3 Image Aesthetic Appeal and sentiment analysis

In [8] system is dealing with the nature of art, beauty with the creation and appreciation of beauty. In this system, we analyze visual content as a machine learning problem, with a peer-rated online image sharing Website as data source. We extract visual features based on the insight that they can differentiate between aesthetically pleasing and displeasing images. By using support vector machines and classification trees we built automated classifiers. The work demonstrates the relationship between emotions which pictures arouse in people, and their low-level content. We have shown significant correlation between various visual properties of photographic images and their aesthetics ratings. By using a community-based database and ratings certain visual features tend to give better discrimination of aesthetic quality than some others. Our SVM-based classifier is robust enough to produce good accuracy using visual features in separating high and low rated photographs.

In [9] we utilize a large dataset of images crawled from Flickr in order to depict links between visual features and sentiment values extracted from the images' textual metadata. We performed an in-depth analysis of the connection between different image features and sentiment on a sample consisting of more than half a million images from the social sharing site Flickr. Our experiments revealed strong and intuitive dependencies between the sentiment values extracted from metadata and visual features based on color histograms and SIFT visual term representations. In our classification experiments, we further confirmed that visual features can, to a certain degree, help predicting the polarity of sentiment. We are known that this work is just one of many steps; in order to make results applicable for real systems, a combination with additional information obtained through advanced text analysis techniques, considering complementary domain knowledge, and focusing on specific problems domains is of high practical importance.

## 3. Crowdsourcing for Visual Summarization

Crowdsourcing is the process of obtaining needed services, ideas, or content by obtaining contributions from a large group of people, and from an online community, rather than from traditional employees or suppliers. It is a relatively upcoming discipline to assure high quality of results.

### 3.1 Crowd Sourcing Experiment

In [10] we study Amazon Mechanical Turk is a crowd sourcing online web service which coordinates the supply and the demand of tasks that require human intelligence to complete. It is an online marketplace where employees called workers are recruited by employers called requesters for the execution of tasks called *HIT*s in exchange for a wage called a reward. Workers and requesters can be linked through an ID provided by Amazon. Employers post HITs which are visible only to workers who meet predefined criteria such as country of residence or accuracy in previously completed tasks. When workers log in the website, they find a list of tasks according to various criteria, including pay per HIT and maximum time allotted for the completion. After completion of task, the requester who supplied that task pays him. The quality of results of crowd sourcing experiment depends on the factors such as e.g., payment amount per HIT, task complexity and worker qualification/reputation.

Here we planned our crowdsourcing task as follows. We recruited 20 different MTurk workers per location for manual creation of reference summaries. As some of them repeated the HIT for the other locations as well, the total number of workers used for the task was 697. In this Task we will give a set of 100 images and ask you to select 10 of them for a “visual summary”. By looking at the 10-image visual summary, workers should obtain the same overall opinion as given by the larger 100-image set. After creation of 10-image summary the worker was asked to sort the selected images in the order of importance and briefly explain reasons for selecting each image using a free text input form. Then perform qualitative analysis of the manually generated visual summaries as well as the criteria for image selection reported by the MTurk workers. The analysis shows that most of them select images that are semantically similar to many other images in the collection, making sure at the same time that as many semantically different images as possible are included in the summary.

### 3.2 Learning to Rank-RANKSVM

In automatic selection of images for the summary we set a target to produce a ranked list of images per location, where the rank position of an image serves as an indicator of its suitability for the visual summary. Here we generate the ranked list in a user-informed fashion as follows:

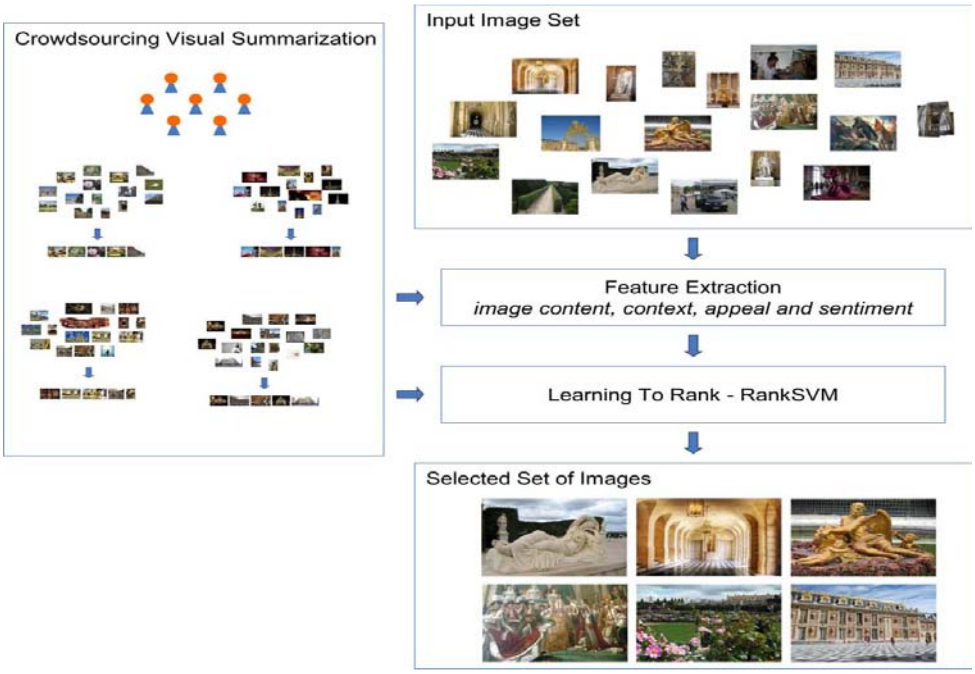
Crowdsourcing Visual Summarization
Input Image Set
Feature Extraction
image content, context, appeal and sentiment
Learning To Rank - RankSVM
Selected Set of Images

**Figure 1:** User-informed approach to image selection for creating visual summaries All images are downloaded from Flickr under CC License [1]

- By selecting the training images from the human-created reference summaries from crowdsourcing experiment
- By learning the ranking function taking the features from image dataset as the input.

We start the training data selection by sorting the images per location based on collection subset consisting of 100 images according to the number of MTurk workers that selected them for their summaries. We choose a set of image preference pairs (i, j) Є P, each consisting of a top ranked and bottom ranked image. Then, to learn the ranking, a well-known RankSVM method is used. In the method originally proposed by Joachims in [12] the RankSVM model is based on minimizing the following objective function.

$$\frac{1}{2}||W||^2 + c \sum_{(i,j)\in P} l(W^T x_i - W^T x_j) \quad \ldots (1)$$

Where
$x_i$ and $x_j$ are the feature vectors representing images
$c$ is regularization parameter
$l$ loss function
W is decision hyperplane normal vector

Due to limitations such as high computational costs associated with training of SVMLight, we make use of a fast RankSVM method described in [11]. This technique uses Newton optimization which does not require computation of difference vectors $x_{i-}$ $x_j$ to significantly reduce the RankSVM training time.

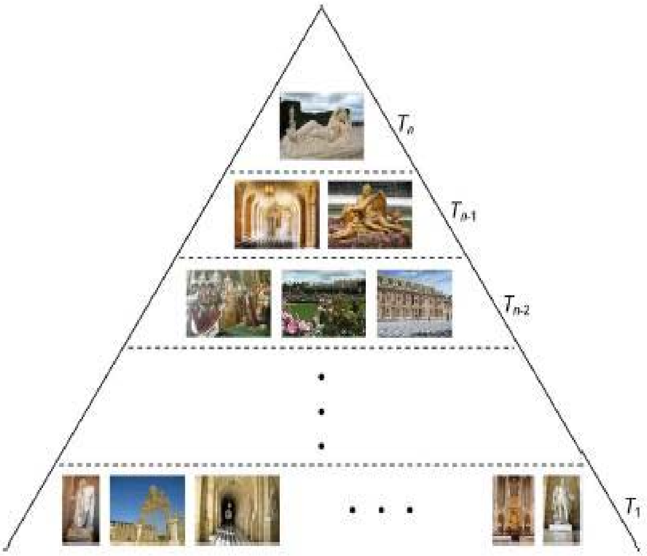

**Figure 2:** Illustration of pyramid structure, where each tier consists of the images appearing in the same number of reference summaries [1]

### 3.3 Summary Set Evaluation: The Pyramid Method

In [13] Summary set evaluation problem in document summaries and machine translations is a low inter-user agreement. Low inter-user agreement is the problem of summary set evaluation in document summaries and machine translation, which makes the evaluation algorithms ROUGH, BLUE and VERT inapplicable. This is the demerit of these evaluation methods.

We present a new pyramid approach for evaluating the suitability of images for the visual summary. As illustrated in Fig.2, every pyramid tier contains the images appearing in the same number of visual summaries. The most frequently selected images are placed in the top level tier, while the bottom level tier is composed of images that were selected by a single Amazon Mechanical Turk worker only. Images that do not present in any of 20 reference summaries generated for a given location are considered unimportant and therefore discarded. We formulate a theory that an optimal set should include all images from the top level tiers and draw the remaining images from the last tier needed to reach a specified set size. As per the pyramid approach an optimal set $\widetilde{R}$ with $N_R$ images would receive the maximum score $d_{max}$ calculated as follows:-

$$d_{max} = \sum_{i=\theta+1}^{n} i \times |T_i| + \theta \times (N_R - \sum_{i=\theta+1}^{n} |T_i|) ,$$

$$\theta = \max_{i} (\sum_{j=i}^{n} |T_j| \geq N_R) \quad \ldots(2)$$

Pyramid scores are reliable, predictive, and conclusive. The pyramid method assigns a score to a summary and based on the score allows the evaluator to evaluate summaries. Pyramid scores are effective in summary evaluation.

## 4. Conclusion

We have studied about how individual select pictures for visual summaries, which was selected from a very large crowd sourcing investigation as the foundation for a novel approach for automatically selecting pictures for visual summarization. The crowd sourcing study revealed natural attributes of pictures that are essential for individuals and also offered people with training data. Our strategy uses characteristics based on these attributes and RankSVM method to create a list of pictures ranked by their suitability for addition in a visual summary. As a result, the selected picture collection can be utilized as a "general purpose" visual summary or as a kick off point in creating overview with specific properties.

## References

[1] Stevan Rudinac, Martha Larson, and Alan Hanjalic, "Learning Crowdsourced User Preferences for Visual Summarization of Image Collections" 1520-9210 © 2013 IEEE

[2] L. S. Kennedy andM. Naaman, "Generating diverse and representative image search results for landmarks," in *Proc. 17th Int. Conf. WorldWide Web, ser. WWW '08*. New York, NY, USA: ACM, 2008, pp.297–306.

[3] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang, "Summarizing tourist destinations by mining user-generated travelogues and photos,"*Comput. Vis. Image Understand.*, vol. 115, no. 3, pp. 352–363, Mar.
2011.

[4] Stevan Rudinac, Alan Hanjalic , "Generating Visual Summaries of geographic Areas Using Community-Contributed Images," IEEE Transactions On Multimedia, Vol. 15, No. 6, 2013

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th Annu.* Meeting on Association for Computational Linguistics, ser. ACL*'02*. Stroudsburg, PA, USA: Association for Computational Linguistics,
2002, pp. 311–318.

[6] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Proc. ACL-04 Workshop Text Summarization Branches *Out*. Barcelona, Spain: Association for Computational Linguistics,2004, pp. 74–81.

[7] Y. Li and B. Merialdo, "VERT: automatic evaluation of video summaries,"in Proc. Int. Conf.Multimedia, ser.MM'10. NewYork,NY,USA: ACM, 2010, pp. 851–854.

[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach", in Proc. 9th Eur. Conf. Computer Vision.Springer 2006, pp. 288-301

[9] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the

social web," in *Proc. Int. Conf.Multimedia, ser. MM '10*. New York, NY, USA: ACM, 2010, pp.715–718.

[10] G.Kazai, "In search of quality in crowdsourcing for search engine evaluation,"in Advances in Information Retrieval, ser. Lecture Notes in *Computer Science*. Berlin, Germany: Springer, 2011, vol. 6611, pp.165–176.

[11] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with SVMS," *Inf. Retriev.*, vol. 13, no. 3, pp. 201–215, Jun. 2010.

[12] T. Joachims, "Optimizing search engines using clickthrough data," in Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and *Data Mining, ser. KDD '02*. New York, NY, USA: ACM, 2002, pp.133–142.

[13] A. Nenkova and R. J. Passonneau, "Evaluating content selection in summarization: The pyramid method," *HLT-NAACL*, pp. 145–152,2004.

[14] 165–176.G.Kazai, "In search of quality in crowdsourcing for search engine evaluation," in *Advances in Information Retrieval, ser. Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2011, vol. 6611, pp. 165–176.

[15] D. Harman and P. Over, "The DUC summarization evaluations," in Proc. 2nd Int. Conf. Human Language Technol. Res., ser. HLT '02, 2002, pp. 44–51, Morgan Kaufmann Publishers

[16] C. Eickhoff and A. de Vries ,"How crowdsourcable is your task?," in *Proc. CSDM '11*, 2011.

## Author Profile

**Ms. Rupali Tanaji Waghmode** has obtained her B.E. in Information Technology from Walchand College of Engineering, Sangli and she is now pursuing her M.E. in Computer Engineering from Zeal Education Society's Dnyanganga College of Engineering and Research, Pune, India

**Prof. Nikita J. Kulkarni** has been working as an Assistant Professor at Zeal Education Society's Dnyanganga College of Engineering and Research, Pune, India